

What is claimed is:

1. A method of clustering genes, comprising:
  - for each condition  $k$  of  $n$  conditions, determining a probability  $p_{ik}$  of a first gene  $g_i$  being in its induced state and a probability  $p_{jk}$  of a second gene  $g_j$  being in its induced state;
  - deriving a contingency table for the first gene  $g_i$  and the second gene  $g_j$  based on the probabilities  $p_{ik}$  and  $p_{jk}$ ;
  - deriving a mutual information  $M$  for the first gene  $g_i$  and the second gene  $g_j$  based on the contingency table; and
  - clustering the first gene  $g_i$  and the second gene  $g_j$  based on the mutual information  $M$  as a metric.
2. The method of claim 1, wherein determining the probability  $p_{ik}$  includes calculating the probability  $p_{ik}$  based on a probability function.
3. The method of claim 1, wherein determining the probability  $p_{ik}$  includes calculating the probability  $p_{ik}$  based on:
$$p_{ik} = g_{\mu_i}(e_{ik} - \theta_i),$$
20 wherein  $g_{\mu_i}(x) = 1/(1 + e^{-\mu_i x})$ ,  $e_{ik}$  represents an expression level of the first gene  $g_i$  under condition  $k$ , and  $\mu_i$  and  $\theta_i$  represent parameters associated with the first gene  $g_i$ .
4. The method of claim 1, wherein determining the probability  $p_{ik}$  includes calculating the probability  $p_{ik}$  based on:
$$p_{ik} = r_{ik}^{\mu_i} / (r_{ik}^{\mu_i} + c_{ik}^{\mu_i} e^{\theta_i \mu_i}),$$
25 wherein  $r_{ik}$  represents an expression level of the first gene  $g_i$  under condition  $k$ ,  $c_{ik}$  represents a control expression level of the first gene  $g_i$  under condition  $k$ , and  $\mu_i$  and  $\theta_i$  represent parameters associated with the first gene  $g_i$ .

5. The method of claim 4, wherein  $\mu_i = 1$ , and  $\theta_i = 0$ .
6. The method of claim 1, wherein deriving the contingency table includes  
5 calculating a 2x2 contingency table  $T_{ij,xy}$  based on:

$$T_{ij,xy} = \begin{pmatrix} \sum_{k=1}^n q_{ik} q_{jk} & \sum_{k=1}^n q_{ik} p_{jk} \\ \sum_{k=1}^n p_{ik} q_{jk} & \sum_{k=1}^n p_{ik} p_{jk} \end{pmatrix},$$

wherein  $q_{ik} = 1 - p_{ik}$ ,  $q_{jk} = 1 - p_{jk}$ ,  $x$  ranges from 0 to 1, and  $y$  ranges from 0 to 1.

7. The method of claim 6, wherein deriving the mutual information  $M$  includes  
10 calculating the mutual information  $M$  based on:

$$M = \sum_{x=0,1} \sum_{y=0,1} P_{ij,xy} \times \log_2 \frac{P_{ij,xy}}{P_{ix} \times P_{jy}},$$

wherein  $P_{ij,xy} = T_{ij,xy}/n$ ,  $P_{ix} = \sum_{y=0,1} T_{ij,xy} / n$ , and  $P_{jy} = \sum_{x=0,1} T_{ij,xy} / n$ .

8. A method of clustering genes, comprising:  
15 for each condition  $k$  of  $n$  conditions, determining a probability  $p_{ik}$  of a first gene  $g_i$  being in its induced state based on a first probability function and a probability  $p_{jk}$  of a second gene  $g_j$  being in its induced state based on a second probability function;  
deriving a contingency table for the first gene  $g_i$  and the second gene  $g_j$  based on the probabilities  $p_{ik}$  and  $p_{jk}$ ;  
20 deriving a mutual information  $M$  for the first gene  $g_i$  and the second gene  $g_j$  based on the contingency table; and  
clustering the first gene  $g_i$  and the second gene  $g_j$  based on the mutual information  $M$  as a metric.

- 25 9. The method of claim 8, wherein at least one of the first probability function and the second probability function corresponds to a sigmoidal probability function.

10. The method of claim 8, wherein deriving the contingency table includes calculating a 2x2 contingency table  $T_{ij,xy}$  based on:

$$T_{ij,xy} = \begin{pmatrix} \sum_{k=1}^n q_{ik} q_{jk} & \sum_{k=1}^n q_{ik} p_{jk} \\ \sum_{k=1}^n p_{ik} q_{jk} & \sum_{k=1}^n p_{ik} p_{jk} \end{pmatrix},$$

5 wherein  $q_{ik} = 1 - p_{ik}$ ,  $q_{jk} = 1 - p_{jk}$ ,  $x$  ranges from 0 to 1, and  $y$  ranges from 0 to 1.

11. A method of deriving a mutual information  $M$  for a first gene  $g_i$  and a second gene  $g_j$ , comprising:

for each condition  $k$  of  $n$  conditions, determining a probability  $p_{ik}$  of the first gene 10  $g_i$  being in its induced state and a probability  $p_{jk}$  of the second gene  $g_j$  being in its induced state;

deriving a 2x2 contingency table  $T_{ij,xy}$  for the first gene  $g_i$  and the second gene  $g_j$  based on the probabilities  $p_{ik}$  and  $p_{jk}$ , wherein  $x$  ranges from 0 to 1, and  $y$  ranges from 0 to 1; and

15 deriving the mutual information  $M$  for the first gene  $g_i$  and the second gene  $g_j$  based on the 2x2 contingency table  $T_{ij,xy}$ .

12. The method of claim 11, wherein determining the probability  $p_{ik}$  includes calculating the probability  $p_{ik}$  based on:

20  $p_{ik} = g_{\mu_i}(e_{ik} - \theta_i),$

wherein  $g_{\mu_i}(x) = 1/(1 + e^{-\mu_i x})$ ,  $e_{ik}$  represents an expression level of the first gene  $g_i$  under condition  $k$ , and  $\mu_i$  and  $\theta_i$  represent parameters associated with the first gene  $g_i$ .

13. The method of claim 11, wherein determining the probability  $p_{ik}$  includes 25 calculating the probability  $p_{ik}$  based on:

$$p_{ik} = r_{ik}^{\mu_i} / (r_{ik}^{\mu_i} + c_{ik}^{\mu_i} e^{\theta_i \mu_i}),$$

wherein  $r_{ik}$  represents an expression level of the first gene  $g_i$  under condition  $k$ ,  $c_{ik}$  represents a control expression level of the first gene  $g_i$  under condition  $k$ , and  $\mu_i$  and  $\theta_i$  represent parameters associated with the first gene  $g_i$ .

5 14. The method of claim 11, wherein deriving the 2x2 contingency table  $T_{ij,xy}$  includes calculating the 2x2 contingency table  $T_{ij,xy}$  based on:

$$T_{ij,xy} = \begin{pmatrix} \sum_{k=1}^n q_{ik} q_{jk} & \sum_{k=1}^n q_{ik} p_{jk} \\ \sum_{k=1}^n p_{ik} q_{jk} & \sum_{k=1}^n p_{ik} p_{jk} \end{pmatrix},$$

wherein  $q_{ik} = 1 - p_{ik}$ , and  $q_{jk} = 1 - p_{jk}$ .

10 15. A method of generating a list of genes, comprising:  
providing a set of gene expression data associated with a plurality of genes under  $n$  conditions;  
selecting a first subset of gene expression data from the set of gene expression data, the first subset of gene expression data being associated with a first gene  $g_i$ ;  
15 selecting a second subset of gene expression data from the set of gene expression data, the second subset of gene expression data being associated with a second gene  $g_j$ ;  
for each condition  $k$  of the  $n$  conditions, determining a probability  $p_{ik}$  of the first gene  $g_i$  being in its induced state based on the first subset of gene expression data;  
for each condition  $k$  of the  $n$  conditions, determining a probability  $p_{jk}$  of the second gene  $g_j$  being in its induced state based on the second subset of gene expression data;  
20 deriving a mutual information  $M$  for the first gene  $g_i$  and the second gene  $g_j$  based on the probabilities  $p_{ik}$  and  $p_{jk}$ ; and  
based on the mutual information  $M$ , generating the list of genes indicating the first  
25 gene  $g_i$  and the second gene  $g_j$ .

16. The method of claim 15, wherein determining the probability  $p_{ik}$  includes calculating the probability  $p_{ik}$  based on:

$$p_{ik} = g_{\mu_i} (e_{ik} - \theta_i),$$

wherein  $g_{\mu_i}(x) = 1/(1 + e^{-\mu_i x})$ ,  $e_{ik}$  represents an expression level of the first gene  $g_i$  under condition  $k$ , and  $\mu_i$  and  $\theta_i$  represent parameters associated with the first gene  $g_i$ .

5 17. The method of claim 15, wherein determining the probability  $p_{ik}$  includes calculating the probability  $p_{ik}$  based on:

$$p_{ik} = r_{ik}^{\mu_i} / (r_{ik}^{\mu_i} + c_{ik}^{\mu_i} e^{\theta_i \mu_i}),$$

wherein  $r_{ik}$  represents an expression level of the first gene  $g_i$  under condition  $k$ ,  $c_{ik}$  represents a control expression level of the first gene  $g_i$  under condition  $k$ , and  $\mu_i$  and  $\theta_i$  represent parameters associated with the first gene  $g_i$ .

10 18. The method of claim 15, further comprising:  
calculating a 2x2 contingency table  $T_{ij,xy}$  based on:

$$T_{ij,xy} = \begin{pmatrix} \sum_{k=1}^n q_{ik} q_{jk} & \sum_{k=1}^n q_{ik} p_{jk} \\ \sum_{k=1}^n p_{ik} q_{jk} & \sum_{k=1}^n p_{ik} p_{jk} \end{pmatrix},$$

15 wherein  $q_{ik} = 1 - p_{ik}$ ,  $q_{jk} = 1 - p_{jk}$ ,  $x$  ranges from 0 to 1, and  $y$  ranges from 0 to 1.

19. The method of claim 18, wherein deriving the mutual information  $M$  includes calculating the mutual information  $M$  based on:

$$M = \sum_{x=0,1} \sum_{y=0,1} P_{ij,xy} \times \log_2 \frac{P_{ij,xy}}{P_{ix} \times P_{jy}},$$

20 wherein  $P_{ij,xy} = T_{ij,xy}/n$ ,  $P_{ix} = \sum_{y=0,1} T_{ij,xy} / n$ , and  $P_{jy} = \sum_{x=0,1} T_{ij,xy} / n$ .

20. A computer-readable medium, comprising:

code to determine, for each condition  $k$  of  $n$  conditions, a probability  $p_{ik}$  of a first gene  $g_i$  being in its induced state and a probability  $p_{jk}$  of a second gene  $g_j$  being in its induced state;

code to derive a contingency table for the first gene  $g_i$  and the second gene  $g_j$

5 based on the probabilities  $p_{ik}$  and  $p_{jk}$ ;

code to derive a mutual information  $M$  for the first gene  $g_i$  and the second gene  $g_j$  based on the contingency table; and

code to cluster the first gene  $g_i$  and the second gene  $g_j$  based on the mutual information  $M$  as a metric.

10

21. The computer-readable medium of claim 20, wherein the code to determine the probability  $p_{ik}$  includes code to calculate the probability  $p_{ik}$  based on a probability function.

15 22. The computer-readable medium of claim 20, wherein the code determine the probability  $p_{ik}$  includes code to calculate the probability  $p_{ik}$  based on:

$$p_{ik} = g_{\mu_i}(e_{ik} - \theta_i),$$

wherein  $g_{\mu_i}(x) = 1/(1 + e^{-\mu_i x})$ ,  $e_{ik}$  represents an expression level of the first gene  $g_i$  under condition  $k$ , and  $\mu_i$  and  $\theta_i$  represent parameters associated with the first gene  $g_i$ .

20

23. The computer-readable medium of claim 20, wherein the code to determine the probability  $p_{ik}$  includes code to calculate the probability  $p_{ik}$  based on:

$$p_{ik} = r_{ik}^{\mu_i} / (r_{ik}^{\mu_i} + c_{ik}^{\mu_i} e^{\theta_i \mu_i}),$$

wherein  $r_{ik}$  represents an expression level of the first gene  $g_i$  under condition  $k$ ,  $c_{ik}$  represents a control expression level of the first gene  $g_i$  under condition  $k$ , and  $\mu_i$  and  $\theta_i$  represent parameters associated with the first gene  $g_i$ .

25

24. The computer-readable medium of claim 23, wherein  $\mu_i = 1$ , and  $\theta_i = 0$ .

25. The computer-readable medium of claim 20, wherein the code to derive the contingency table includes code to calculate a 2x2 contingency table  $T_{ij,xy}$  based on:

$$T_{ij,xy} = \begin{pmatrix} \sum_{k=1}^n q_{ik} q_{jk} & \sum_{k=1}^n q_{ik} p_{jk} \\ \sum_{k=1}^n p_{ik} q_{jk} & \sum_{k=1}^n p_{ik} p_{jk} \end{pmatrix},$$

wherein  $q_{ik} = 1 - p_{ik}$ ,  $q_{jk} = 1 - p_{jk}$ ,  $x$  ranges from 0 to 1, and  $y$  ranges from 0 to 1.

5

26. The computer-readable medium of claim 20, further comprising:  
code to receive hybridization data of a sample nucleic acid sequence and nucleic acid probes, wherein the code to determine the probability  $p_{ik}$  includes code to calculate the probability  $p_{ik}$  based on the hybridization data.

10

27. The computer-readable medium of claim 20, further comprising:  
code to interface with a hypertext transfer protocol server.

15

28. The computer-readable medium of claim 20, further comprising:  
code to generate a hypertext markup language document indicating the first gene  $g_i$  and the second gene  $g_j$ .